Scalable Data Extraction Techniques for Transforming Electronic Documents into Queriable Archives

Abstract of the Disclosure

A method for extracting an attribute occurrence from template generated semi-structured document comprising multi-attribute data records comprises identifying a first set of attribute occurrences in the template generated semi-structured document using an ontology. The method further comprises determining a boundary of each multi-attribute data record in the template generated semi-structured document, learning a pattern for an attribute corresponding to an identified attribute occurrence of the first set in the template generated semi-structured document, and applying the pattern within the boundary of each multi-attribute data record in the template generated semi-structured document to extract a second set of attribute occurrences.